

Bipartite Community Structure of eQTLs

John Platig^{*,1,2}, Peter Castaldi^{3,4,5,6}, Dawn DeMeo^{3,5,6}, and John Quackenbush^{1,2,3}

¹Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
Boston, MA

²Department of Biostatistics, Harvard Chan School of Public Health, Boston, MA

³Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA

⁴Division of General Medicine, Brigham and Women's Hospital, Boston, MA

⁵Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital,
Boston, MA

⁶Harvard Medical School, Boston, MA

September 10, 2015

1 Abstract

Keywords: Genomics, GWAS, eQTL, COPD, Bipartite Networks, Community Structure

Genome Wide Association Studies (GWAS) and eQTL analyses have produced a large and growing number of genetic associations linked to a wide range of human phenotypes. As of 2013, there were more than 11,000 SNPs associated with a trait as reported in the NHGRI GWAS Catalog. However, interpreting the functional roles played by these SNPs remains a challenge. Here we describe an approach that uses the inherent bipartite structure of eQTL networks to place SNPs into a functional context.

Using genotyping and gene expression data from 163 lung tissue samples in a study of Chronic Obstructive Pulmonary Disease (COPD) we calculated eQTL associations between SNPs and genes and cast significant associations ($\text{FDR} < 0.1$) as links in a bipartite network. To our surprise, we discovered that the highly-connected “hub” SNPs within the network were devoid of disease-associations. However, within the network we identified 35 highly modular communities, which comprise groups of SNPs associated with groups of genes; 13 of these communities were significantly enriched for distinct biological functions ($P < 5 \times 10^{-4}$) including COPD-related functions. Further, we found that GWAS-significant SNPs were enriched at the cores of these communities, including previously identified GWAS associations for COPD, asthma, and pulmonary function,

*jplatig@jimmy.harvard.edu

among others. These results speak to our intuition: rather than single SNPs influencing single genes, we see groups of SNPs associated with the expression of families of functionally related genes and that disease SNPs are associated with the perturbation of those functions. These methods are not limited in their application to COPD and can be used in the analysis of a wide variety of disease processes and other phenotypic traits.

2 Introduction

Genome Wide Association Studies (GWAS) have created new opportunities to understand the genetic factors that influence complex traits. Excepting highly-penetrant Mendelian disorders, the majority of genetic associations seem to be driven by many factors, each of which has a relatively small effect. In a recent study [24], 697 SNPs were associated with height in humans at genome-wide significance, yet these SNPs were able to explain only $\sim 20\%$ of height variability; $\sim 9,500$ SNPs were needed to raise that to $\sim 29\%$. In addition, $\sim 95\%$ of GWAS variants map to non-coding regions [3], complicating biological interpretation of their functional impact.

To bridge the functional gap between genetic variant and complex trait, expression Quantitative Trait Locus (eQTL) analysis associates SNP genotype with gene expression levels. Most eQTL analyses have focused on *cis*-SNPs—those near the Transcriptional Start Site (TSS) of the gene in the association test. Recent computational developments [22] and work demonstrating the impact and replicability of *trans*-eQTLs [23, 16] have increased interest in identifying and understanding the role played by *trans*-acting SNPs.

However, new methods are needed to elucidate the potential functional impact of the thousands of GWAS and eQTL SNPs that can be detected in a single study. Here we present a complex networks method (Fig. 1) that incorporates both *cis*- and *trans*- associations to identify groups of SNPs that are linked to groups of genes and systematically interrogate their biological functions. We then validate this approach using genotyping and gene expression data from 163 lung tissue samples in a study of Chronic Obstructive Pulmonary Disease (COPD) by the Lung Genomics Research Consortium (LGRC).

eQTL Networks

We used the `MatruxeQTL` package in R to calculate all *cis*- and *trans*-eQTLs, considering only autosomal SNPs, using age, sex, and pack-years as covariates (see Supplementary Methods). The *cis*- and *trans*- associations were run separately, with an FDR threshold of 10%. This analysis identified 32,053 *cis*-eQTLs and 39,107 *trans*-eQTLs. Quantile-quantile plots for both *cis*- and *trans*- are shown in Supplementary Figure S1. In total, 71,160 statistically significant associations were detected between 54,475 SNPs and 7,091 genes.

We represented these associations as a bipartite network consisting of two classes of nodes—SNPs and genes—with edges from SNPs to the genes with which they are significantly associated

based on the eQTL FDR cut-off. The network had a Giant Connected Component (GCC) with 44,872 links, 29,907 SNPs, and 3,390 genes. As a network diagnostic, we plotted the distributions of edges per SNP—called the SNP degree—and edges per gene (Fig. 2) for the SNPs and genes in the GCC. The degree distributions for both the SNPs and genes are broad-tailed, implying potential power-law behavior. To test this, we fit each degree distribution to a power law, and determined the goodness of fit using the method described in [11] (see Supplementary Methods). The probability the SNP degree follows a power-law distribution is very low, $P_{pl} \approx 0$, and the gene degree distribution (Fig. 2b) is also unlikely to be power-law distributed ($P_{pl} < 0.1$) even though there are multiple network hubs, shown in the tail of the distribution in Figure 2b.

It is often cited in complex networks literature that the hubs, those nodes in the network that are most highly connected, represent critical elements whose removal can disrupt the entire network [2, 18]. As a result, one widely-held belief about biological networks is that disease-related elements should be over-represented among the network hubs [4]. To test the hypothesis that disease-associated SNPs are concentrated in the hubs, we projected GWAS-identified SNPs associated with a wide range of diseases and phenotypes onto the SNP degree distribution (Fig. 3). We used the *gwascat* package [8] in R to download GWAS SNPs annotated in the NHGRI GWAS catalog; 259 of those SNPs mapped to the GCC of the eQTL network. To our surprise, the network hubs—the right tail of Figure 3—were devoid of disease-associated SNPs which were instead scattered through the upper left half of the degree distribution. While the SNPs associated with a single gene are easier to interpret, the concentration of disease-associated SNPs in the middle of the distribution prompted us to look at other features of the network and its structure.

Community Structure Analysis

Many real-world networks have a complex structure consisting of “communities” of nodes [17]. These communities are often defined as a group of network nodes that are more likely to be connected to other nodes within their community than they are to those outside of the community. A widely used measure of community structure is the modularity, which can be interpreted as an enrichment for links within communities minus an expected enrichment given the network degree distribution [20].

To partition the nodes from the eQTL network into communities—each of which contains both SNPs and genes—we maximized the bipartite modularity [5]. As recursive cluster identification and optimization can be computationally slow, we calculated an initial community structure assignment on the weighted, gene-space projection, using a fast uni-partite modularity maximization algorithm [6] available in the R *igraph* package [14], then iteratively converged ($\Delta Q < 10^{-4}$) on a community structure corresponding to a maximum bipartite modularity.

The bipartite modularity is defined in Eq. (1), where m is the number of links in the network, \tilde{A}_{ij} is the upper right block of the network adjacency matrix (a binary matrix where a 1 represents a connection between a SNP and a gene and 0 otherwise), k_i is the degree of SNP i , d_j is the degree of gene j , and C_i , C_j the community indices of SNP i and gene j , respectively.

$$Q = \frac{1}{m} \sum_{i,j} \left(\tilde{A}_{ij} - \frac{k_i d_j}{m} \right) \delta(C_i, C_j) \quad (1)$$

This analysis identified 35 highly modular communities in the LGRC data ($Q = 0.766$; Fig. 4). The density of these communities can be seen in Figure 4. In Figure 4b, there is visible enrichment for links within each community (colored links) compared to links between different communities (black links). These communities represent groups of SNPs and genes that are highly connected to each other, suggesting that groups of genes may be jointly moderated by groups of SNPs that together represent specific biological processes.

To investigate this hypothesis, we tested each community for GO term enrichment using Fisher’s Exact Test (available in the R package *GOstats* [15]) and found 13 of the 35 communities contained genes enriched for specific Gene Ontology terms ($P < 5 \times 10^{-4}$; overlap > 4), encompassing a broad collection of cellular functions that are not generally associated with COPD. Indeed, this is what one might expect as the genetic background of an individual should have an effect not only on disease-specific processes, but more globally on the physiology of his or her individual cells. A number of communities do, however, show enrichment for biological processes that are known to be involved in COPD, including genes previously associated with the disease.

For example, Community 18 (see Fig. 4) was enriched for chromatin and nucleosome assembly/organization and includes members of the HIST1H gene superfamily. Community 30 (see Fig. 4) included GO term enrichment for function related to the HLA gene family, including T cell function and immune response; autoimmunity has been suggested as a potential contributor to COPD pathogenesis [1]. This community also contains PSORS1C1, which has been previously implicated in COPD [21].

Another of the genes in Community 30, *AGER*, has been implicated in COPD [9] and encodes sRAGE, a biomarker for emphysema. Its expression is negatively associated via eQTL analysis ($\beta = -0.3$) with rs6924102. This SNP has been observed to be an eQTL in a large blood eQTL dataset for a number of neighboring genes [23], but it has not previously been described as an eQTL for *AGER*. This SNP lies in a region containing a DNase peak in cell lines analyzed by ENCODE [12] (indicating it sits in a region of open chromatin) and there is evidence of POLR2A binding from ChIP-Seq data in the GM12878 cell line as reported by ENCODE (<http://regulomedb.org/snp/chr6/32811382>). This suggests that rs6924102 may inhibit the expression of *AGER* through disruption of RNA Polymerase II binding and subsequent mRNA synthesis. This SNP is located ~ 700 KB from the well-studied non-synonymous *AGER* SNP, rs2070600.

Examining Figure 4a, it is evident that within each community there are local hubs that are highly connected to the genes within that community. To identify these local hubs, we defined a “core score” that estimates importance of a SNP in the structure of its community. For SNP i in community h , its core score, Q_{ih} , Eq. (2), is the fraction of the modularity of community h , Q_h , Eq. (3), contributed by SNP i . This allows for comparison of SNPs from different communities, as

each community does not have the same modularity, Q_h .

$$Q_{ih} = \frac{\frac{1}{m} \sum_j \left(\tilde{A}_{ij} - \frac{k_i d_j}{m} \right) \delta(C_i, h) \delta(C_j, h)}{Q_h} \quad (2)$$

$$Q_h = \frac{1}{m} \sum_{i,j} \left(\tilde{A}_{ij} - \frac{k_i d_j}{m} \right) \delta(C_i, h) \delta(C_j, h) \quad (3)$$

If one views disease as the disruption of a process leading to cellular or organismal dysfunction, one natural hypothesis is that SNPs with the greatest potential to disrupt cellular processes might be enriched for disease association. To test this we used both the Wilcoxon rank-sum and Kolmogorov-Smirnov (KS) tests to assay whether the 259 NHGRI GWAS-annotated SNPs in the GCC were more likely to have high Q_{ih} scores. For both tests, the distribution of Q_{ih} scores for GWAS-associated SNPs were compared to the distribution of non-GWAS SNP scores.

To obtain an empirical p-value for these tests, we permuted the GWAS/non-GWAS labels and recalculated the KS and Wilcoxon tests 10^5 times. Histograms of the randomized labels are shown in Figure 5 and (Supplementary Figure S2). The red dot in the histogram represents the test score with the true labeling. Both tests had highly significant permutation p-values, with $P < 10^{-5}$ for the KS and Wilcoxon tests, indicating that GWAS SNPs were over-represented among SNPs with high core scores. Furthermore, the median core score for the GWAS SNPs was 1.69 times higher than the median core score for the non-GWAS SNPs. Thus, while global hubs are depleted for GWAS associations with disease, local hubs are significantly enriched for disease associations.

Discussion

Genome-wide association studies have searched for genomic variants that influence complex traits, including the development and progression of disease. However, the number of highly-penetrant Mendelian variants that have been found is surprisingly small, with most disease-associated SNPs having a weak phenotypic effect. GWAS studies have also identified many SNPs that do not alter protein coding and have found significant loci that are shared in common across multiple diseases. This body of evidence suggests that in most instances it is not a single genetic variant that leads to disease, but many variants of smaller effect that together can disrupt cellular processes that lead to disease phenotypes. The challenge has been to find these variants of small effect and to place them into a coherent biological context.

We chose to address this problem by analyzing the link between genetic variants and the most immediate phenotypic measure, gene expression. In doing so, we chose not to focus solely on *cis*-acting SNPs, but also to consider *trans*-acting variants. Our motivation was, in part, to try to understand SNPs found through GWAS studies to be associated with phenotypes, but that could not be immediately placed into a functional context. After performing a genome-wide *cis*- and *trans*- eQTL analysis, we identified a large number of many-to-many associations: single SNPs associated with many genes as well as single genes that were significantly associated with many

SNPs. To represent those associations, we constructed a bipartite network, one that contains two types of nodes—SNPs and genes—with edges connecting SNPs to the genes with which they were significantly associated. Our analysis of that network led to a number of observations that independently speak to our intuition about disease and the genetic factors that control it.

First is the observation that the highly connected SNPs, the global hubs in the network, are devoid of variants that have been identified as being disease-associated in the hundreds of studies collected in the NHGRI GWAS catalog. While initially surprising, further consideration suggests that this may be the result of negative selection. Since a true hub SNP influences genes across the genome that are involved in many biological processes, highly deleterious variants are likely to significantly disrupt cellular function. In fact, this is the expected impact of a hub—its disruption should lead to the catastrophic collapse of the network. And so, deleterious SNPs that are network hubs are likely to be lethal or highly debilitating and therefore strongly selected against and quickly swept from the genome.

Second, we found that SNPs and their target genes form highly connected communities that are enriched for specific biological functions. This too speaks to our intuition and to the evidence about polygenic traits that has accumulated over time. They are not the result of a single SNP that regulates a single gene, but a family of SNPs that together help mediate a group of functionally-related genes.

Third, the enrichment for GWAS disease associations among the high Q_{ih} SNPs has a very simple and intuitive interpretation. The SNPs that are most significantly connected within a particular functionally-related group are those most likely to disrupt that process and therefore be discovered in GWAS analysis. After all, diseases do not develop because the cell’s entire functionality collapses, but because specific processes within the cell are disrupted.

What our analysis provides is a new way of exploring *cis*- and *trans*- eQTL analysis and GWAS. What one must do is to consider not only the local effects of genetic variants, but also the complex network of genetic interactions that help regulate phenotypes, including gene expression.

It also suggests a new way of filtering genes for inclusion in GWAS analysis. Since many disease-associated SNPs appear to be either *cis*-acting or those which are central to functionally-defined communities, one should focus on those SNPs most likely perturb specific biological processes rather than considering the entirety of SNPs in the genome.

As a way of further assessing the link between GWAS significance and functional perturbation in COPD, we calculated a GWAS-FDR for all SNPs in the GCC of our network that had a reported p-value from a recent GWAS and meta-analysis of COPD [10] (See Supplemental Methods). There were 34 SNPs with an FDR < 0.05, and 32 of the 34 had evidence of functional impact according to RegulomeDB [7], with 16 SNPs identified as likely to affect transcription factor binding and linked to expression (See Fig. 6). These 34 SNPs mapped to 4 different communities including Community 30, which contains other COPD-associated SNPs and genes, and is enriched for GO terms describing T cell function and immune response. One of the SNPs in this community likely to affect binding is rs9268528, which is linked by our network to HLA-DRA, HLA-DRB4, and HLA-

DRB5; the *cis*-eQTL associations between rs9268528 and both HLA-DRA and HLA-DRB5 have been previously observed in lymphoblastoid cells [19]. All three HLA genes lie in Community 30 and contribute to the community’s enrichment for T cell receptor signaling pathway (GO:0050852) [13].

To determine the network influence of these 34 SNPs, we compared their core score, Q_{ih} (see Section 2 and Eq. (2)) to the core scores of SNPs with a GWAS-FDR ≥ 0.05 (See Fig. 7). The median Q_{ih} value for the 34 GWAS-FDR significant SNPs was 46.7 times higher than the median for SNPs with an FDR ≥ 0.05 .

One might note that this analysis was carried out using data on genetic variation and gene expression from the LGRC representing COPD and control lung tissue and question both the generalizability of the results and the use of GWAS-associated disease SNPs from many diseases in the analysis. While these are potentially legitimate concerns, many of the community-based processes we find are not specific to COPD or to the lung but instead are active in nearly all human cell types.

Although one might expect some processes to change in different disease states, the impact of common variants and the structure of the network is likely to be highly similar. Consequently, although there may be some SNPs whose impact is disease and tissue specific, many are likely to be independent of disease state. This suggests that it may be useful to develop eQTL networks across disease states and tissue types and to explore changes in the overall network and community structure across and between phenotypes due to rare variants and tissue-specific expression.

Validating individual associations in the eQTL network is a difficult challenge. Most eQTL studies limit their validation efforts to downstream effects of high-confidence *cis*-acting eQTLs. The bipartite network presented here captures not only these strong *cis*-eQTLs but also the weak effects of many more *cis*- and *trans*-acting SNPs. So the likelihood that any individual association can be easily validated may not be that great, as it is likely to be of small phenotypic effect and important in only a subset of individuals. However, this is not the point. What is important for the phenotype is not any single SNP-gene association, but the “mesoscale” organization of genes and SNPs represented by the communities in the network. We believe this intermediate structure better reflects the aggregation of weak genetic effects that contribute to late-onset complex diseases. What we hope to have demonstrated in this manuscript is that the higher order structure, which was not an input to the model, provides insight into a number of aspects of the genetics of polygenic traits consistent with our understanding of how these traits manifest themselves.

Acknowledgements

This work was supported by grants 1P01HL105339 (PC, DD, JP, JQ) and 1R01HL111759 (DD, JP, JQ) from the US National Heart Lung Blood Institute of the National Institutes of Health and grant 1R01AI099204 from the US National Institute of Allergy and Infectious Disease of the National Institutes of Health (JP, JQ). PC was also supported by grants from the US National Heart, Lung,

and Blood Institute (K08 HL102265 and R01 HL124233). Primary data used in this analysis was generated by the Lung Genomics Research Consortium (RC2 HL101715). Previously published COPD GWAS results were supported by COPDGene(R01 HL089856 and R01 HL089897). We thank Drs. Michael Cho and Ed Silverman for their insight and assistance with the COPD GWAS data.

References

- [1] A Agust, W MacNee, K Donaldson, and M Cosio. Hypothesis: Does copd have an autoimmune component? *Thorax*, 58(10):832–834, 2003.
- [2] Réka Albert, Hawoong Jeong, and Albert-László Barabási. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, 2000.
- [3] Kristin G Ardlie, David S Deluca, Ayellet V Segrè, Timothy J Sullivan, Taylor R Young, Ellen T Gelfand, Casandra A Trowbridge, Julian B Maller, Taru Tukiainen, Monkol Lek, et al. The genotype-tissue expression (gtex) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015.
- [4] Albert-László Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.
- [5] Michael J Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007.
- [6] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.
- [7] Alan P. Boyle, Eurie L. Hong, Manoj Hariharan, Yong Cheng, Marc A. Schaub, Maya Kasowski, Konrad J. Karczewski, Julie Park, Benjamin C. Hitz, Shuai Weng, J. Michael Cherry, and Michael Snyder. Annotation of functional variation in personal genomes using regulomedb. *Genome Research*, 22(9):1790–1797, 2012.
- [8] VJ Carey. *gwascat: representing and modeling data in the NHGRI GWAS catalog*. R package version 1.8.0.
- [9] Donavan T Cheng, Deog Kyeom Kim, Debra A Cockayne, Anton Belousov, Hans Bitter, Michael H Cho, Annelyse Duvoix, Lisa D Edwards, David A Lomas, Bruce E Miller, et al. Systemic soluble receptor for advanced glycation endproducts is a biomarker of emphysema and associated with ager genetic variants in patients with chronic obstructive pulmonary disease. *American journal of respiratory and critical care medicine*, 188(8):948–957, 2013.

- [10] Michael H Cho, Merry-Lynn N McDonald, Xiaobo Zhou, Manuel Mattheisen, Peter J Castaldi, Craig P Hersh, Dawn L DeMeo, Jody S Sylvia, John Ziniti, Nan M Laird, et al. Risk loci for chronic obstructive pulmonary disease: a genome-wide association study and meta-analysis. *The Lancet Respiratory Medicine*, 2(3):214–225, 2014.
- [11] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
- [12] ENCODE Project Consortium et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.
- [13] The Gene Ontology Consortium. Gene ontology consortium: going forward. *Nucleic Acids Research*, 43(D1):D1049–D1056, 2015.
- [14] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.
- [15] S Falcon and R Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–8, 2007.
- [16] Rudolf SN Fehrmann, Ritsert C Jansen, Jan H Veldink, Harm-Jan Westra, Danny Arends, Marc Jan Bonder, Jingyuan Fu, Patrick Deelen, Harry JM Groen, Asia Smolonska, et al. Trans-eqtls reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the hla. *PLoS genetics*, 7(8):e1002197, 2011.
- [17] Michelle Girvan and Mark EJ Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826, 2002.
- [18] Hawoong Jeong, Bálint Tombor, Réka Albert, Zoltan N Oltvai, and A-L Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804):651–654, 2000.
- [19] Stephen B Montgomery, Micha Sammeth, Maria Gutierrez-Arcelus, Radoslaw P Lach, Catherine Ingle, James Nisbett, Roderic Guigo, and Emmanouil T Dermitzakis. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature*, 464(7289):773–777, 2010.
- [20] Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.
- [21] Weiliang Qiu, Michael H Cho, John H Riley, Wayne H Anderson, Dave Singh, Per Bakke, Amund Gulsvik, Augusto A Litonjua, David A Lomas, James D Crapo, et al. Genetics of sputum gene expression in chronic obstructive pulmonary disease. *PLoS One*, 6(9):e24395, 2011.

- [22] Andrey A Shabalín. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.
- [23] Harm-Jan Westra, Marjolein J Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W Christiansen, Benjamin P Fairfax, Katharina Schramm, Joseph E Powell, et al. Systematic identification of trans eqtls as putative drivers of known disease associations. *Nature genetics*, 45(10):1238–1243, 2013.
- [24] Andrew R Wood, Tõnu Esko, Jian Yang, Sailaja Vedantam, Tune H Pers, Stefan Gustafsson, Audrey Y Chu, Karol Estrada, Jian’an Luan, Zoltán Kutalik, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nature genetics*, 46(11):1173–1186, 2014.

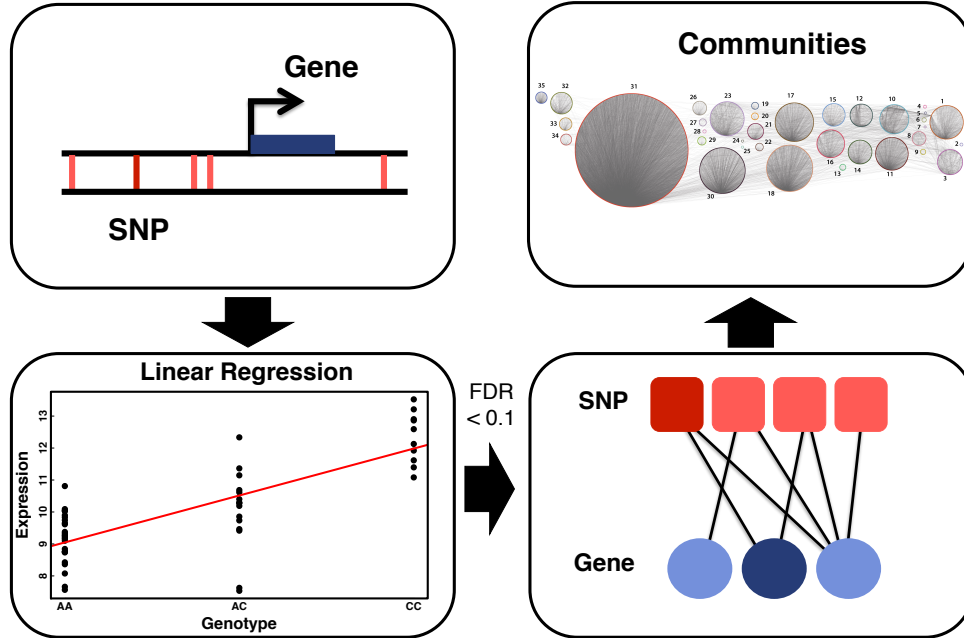


Figure 1: Summary of the Network Method: All possible SNP-Gene pairs measured in the LGRC data set are considered in the eQTL analysis. Those SNP-gene pairs that are significantly associated ($\text{FDR} < 0.1$) are included in the bipartite network. Communities were detected in the network using a bipartite modularity maximization approach, producing 35 highly modular communities, 13 of which are enriched for various GO terms. GWAS-associated SNPs are much more likely to lie at the cores of these communities.

3 Supplemental Methods

We began by downloading gene expression data from the LGRC web portal (<https://www.lung-genomics.org/download/>) representing data from COPD-case and control samples generated by the Lung Genomics Research Consortium (LGRC). This included GCRMA-normalized gene expression data

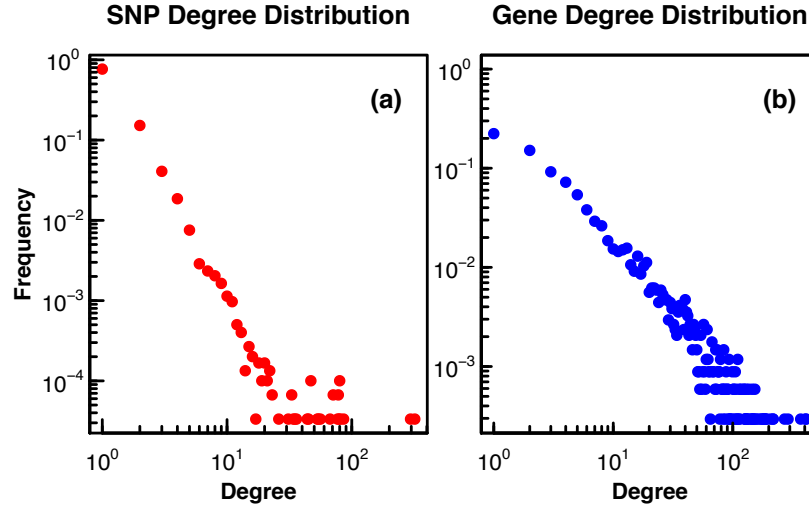


Figure 2: Plot of the degree distribution, showing the frequency of node degree plotted on a log-log scale, for SNPs (left) and genes (right) in the giant connected component of the bipartite eQTL network.

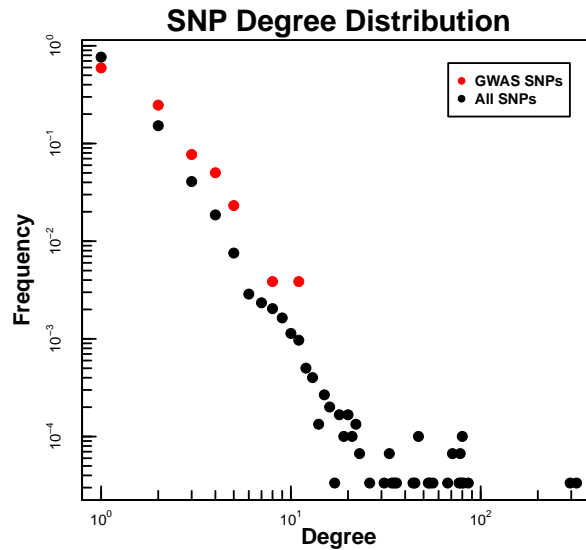


Figure 3: Degree distributions for NHGRI-GWAS (red) and non-GWAS (black) SNPs. NHGRI-GWAS SNPs tend not to be network “hubs,” which are located in the far-right tail of the distribution. The highest degree NHGRI-GWAS SNP was connected to 11 genes.

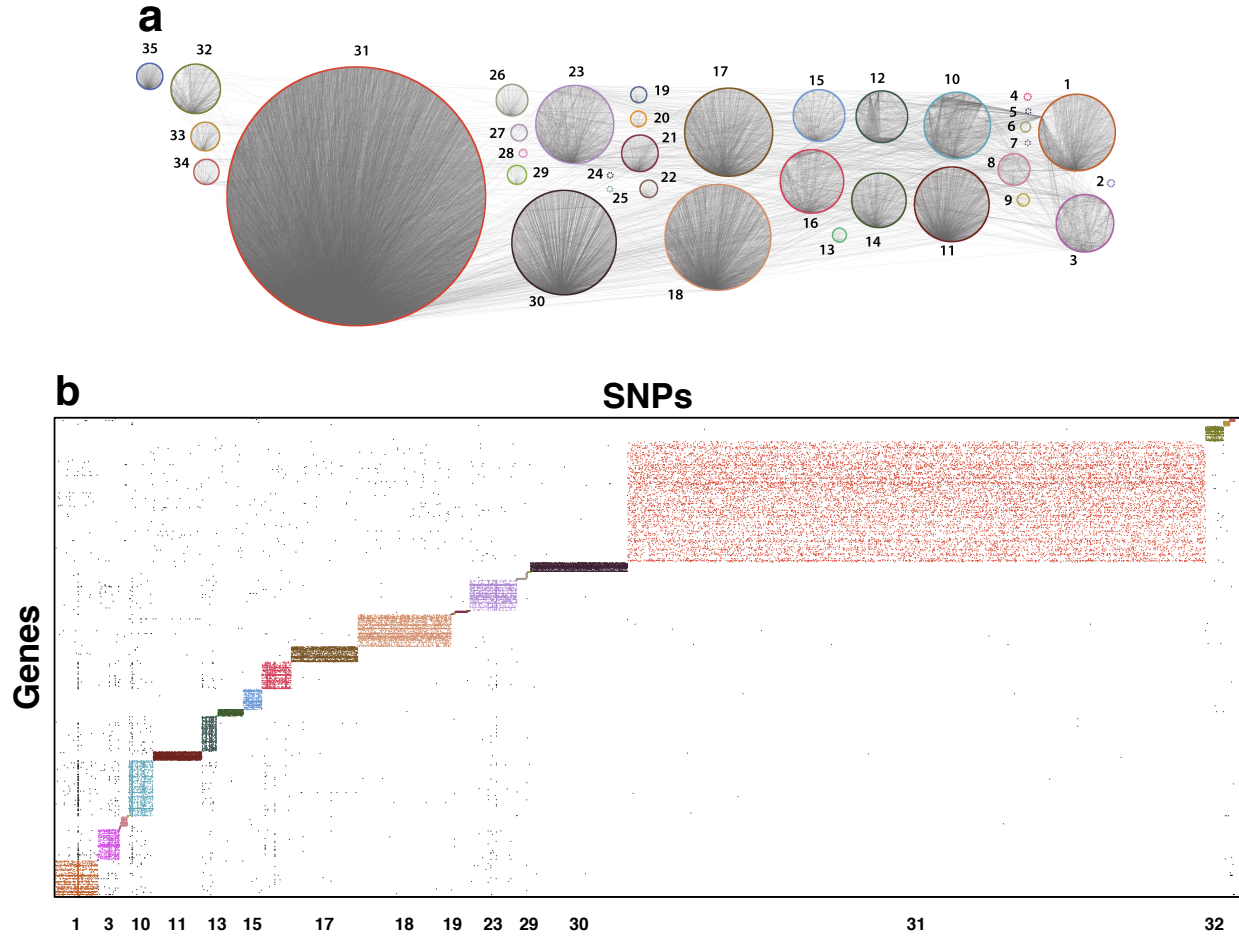


Figure 4: **(a)** Plot of the communities within the bipartite eQTL network. The nodes (genes and SNPs) in each community form a ring, with the link density within each ring visibly darker than links between communities. **(b)** Links within communities (colored points) are shown along the diagonal, with links that go between communities in black. Community IDs are plotted along the x -axis.

obtained using Agilent-014850 Whole Human Genome 4x44K and Agilent-028004 SurePrint G3 Human GE 8x60K Microarrays. We then obtained matching genotyping data (dbGAP accession phs000624.v1.p1) collected using the Illumina Infinium HD Assays with Human Omni 1Quad and Human Omni 2.5 Quad arrays. All subjects were reported to be of Caucasian descent and were selected based on a variety of parameters including clinical measures associated with diagnosis. Samples that did not meet standards for lack of relatedness as measured using Identity by Descent (IBD) and inbreeding coefficient, F , were excluded. Those samples with discordance between reported and genetic sex were not included. Samples missing more than 10% percent of genotyped SNPs were also removed. SNPs with minor allele frequency (MAF) < 0.05 or Hardy Weinberg Equilibrium < 0.001 were removed. After all quality controls, 163 samples remained. The COPD GWAS data from a meta-analysis of COPDGene non-Hispanic whites and African-Americans, ECLIPSE,

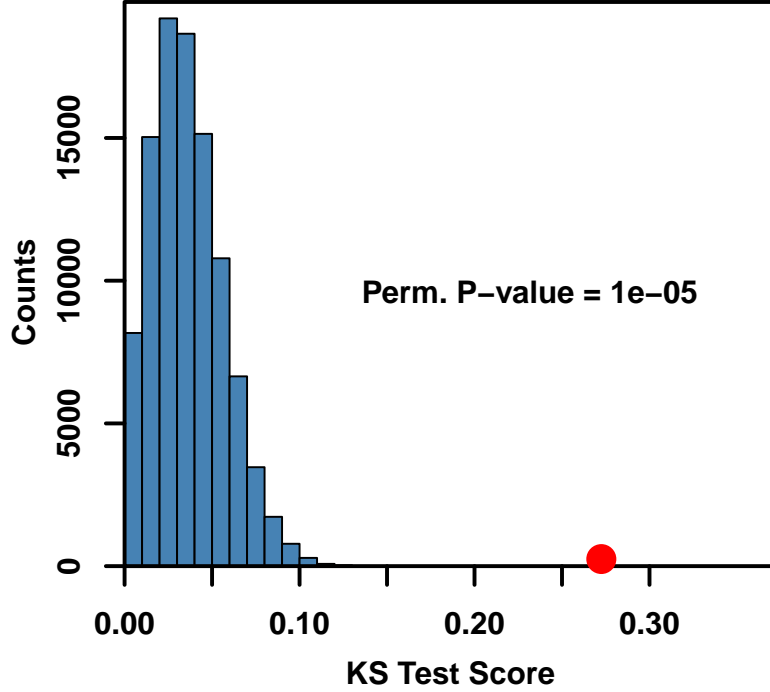


Figure 5: Histogram of Kolmogorov-Smirnov test statistics comparing the distribution of Q_{ih} scores for sets of randomly relabeled NHGRI-GWAS/non-GWAS SNPs. The KS test statistic for the true labeling is in red. The permutation p-value associated with the KS test is $P < 10^{-5}$ given 10^5 permutations.

GenKOLS, and NETT/NAS studies was obtained from the authors of [10].

3.1 Power-law Fitting

For each empirical degree distribution, we fit the two parameters for a power-law: the minimum degree at which the power-law behavior starts, d_{min} , and the exponent, α . A Kolmogorov-Smirnov test was then used to estimate the goodness of fit between 5,000 randomly generated power-law distributed synthetic data sets given d_{min} and α and their corresponding power-law fit. The probability, P_{pl} , that a degree distribution follows a power-law with d_{min} and α is then the fraction of times a synthetic data set has a KS statistic larger than that of the true test. For both the SNP and gene degree distributions, P_{pl} was calculated using the 5,000 goodness of fit values (code for the parameter estimation, goodness of fit and probability estimation was obtained from <http://tuvalu.santafe.edu/~aaronc/powerlaws/>).

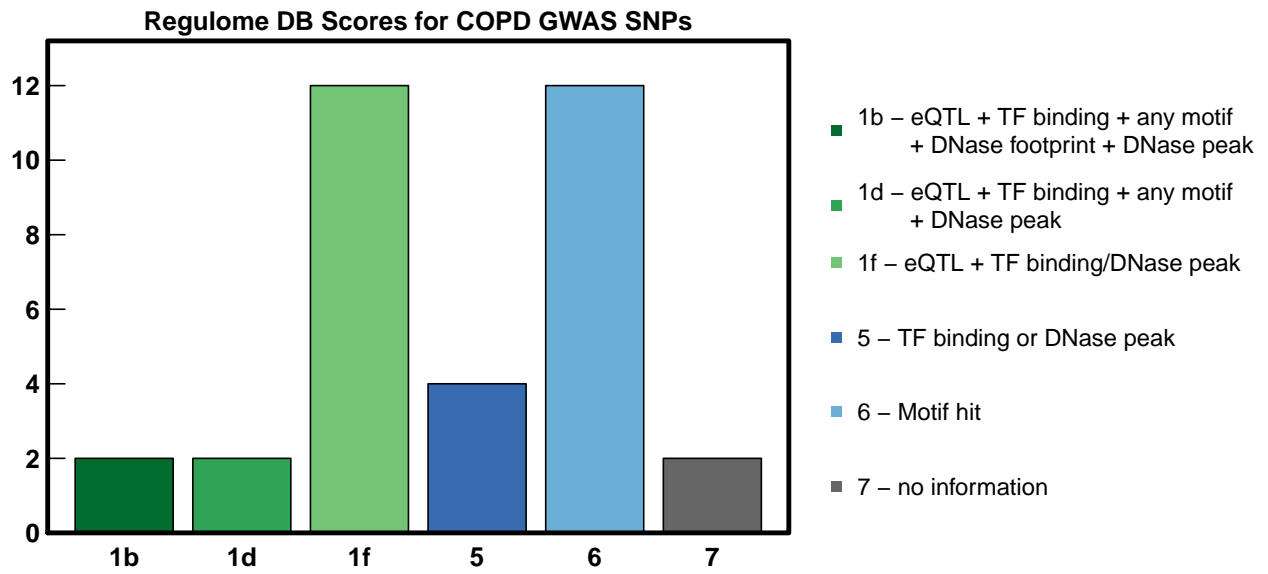


Figure 6: Of the 34 SNPs that are eQTLs in the GCC of the LGRC network and also associated with COPD ($FDR < 0.05$), 16 are likely to affect transcription factor (TF) binding and linked to the expression of a target gene (a score of 1b, d, or f), 4 have evidence of TF binding or a DNase peak (a score of 5), and 12 are located in a motif hit (a score of 6) according to RegulomeDB [7].

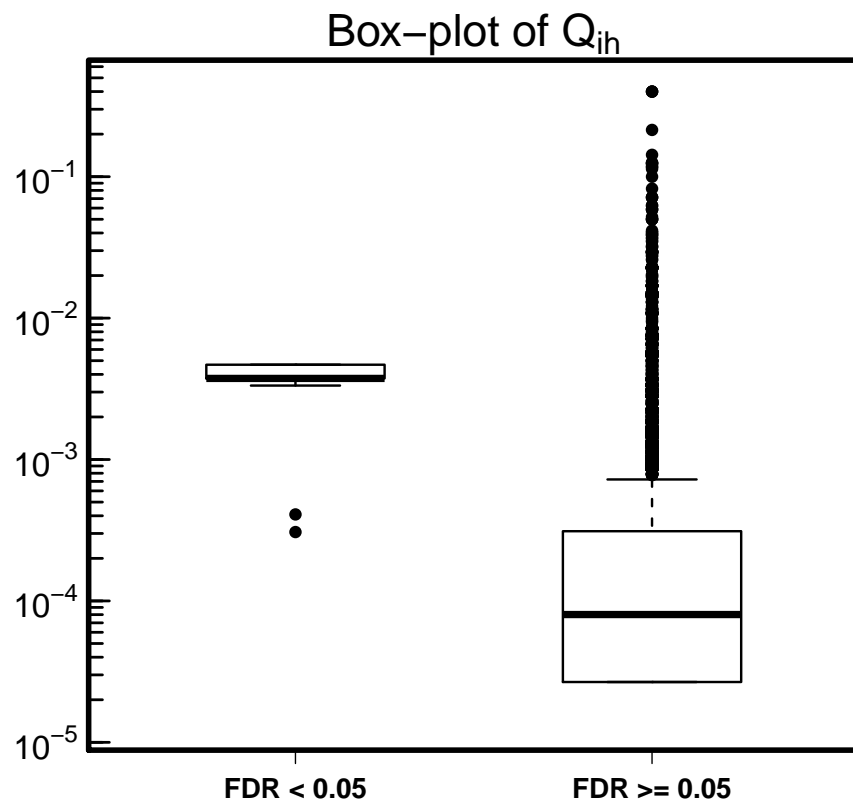


Figure 7: The median core scores for the 34 FDR-significant COPD GWAS SNPs (FDR < 0.05, left) is 46.7 times higher than the median core score for the non-significant SNPs (FDR \geq 0.05, right).

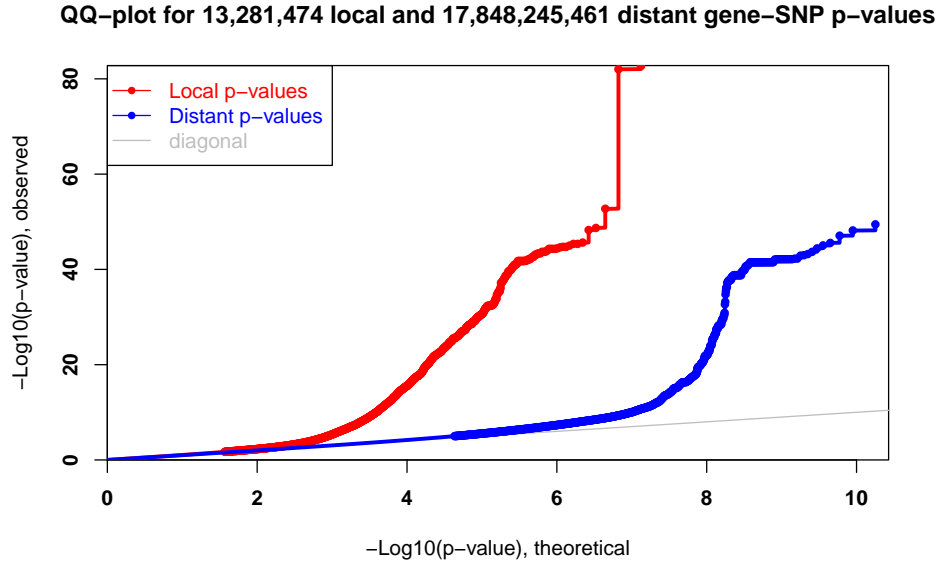


Figure S1: Quantile-quantile plot for *cis*- and *trans*-eQTL associations.

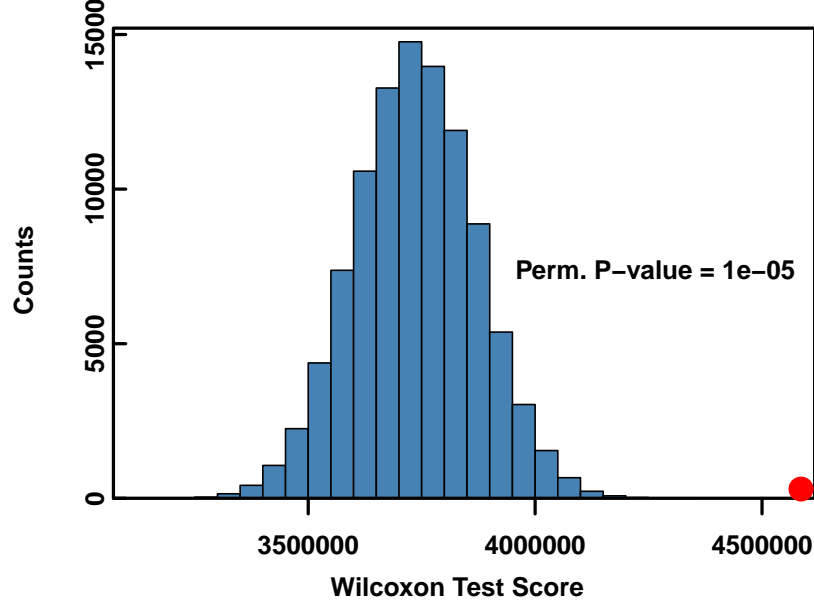


Figure S2: Histogram of Wilcoxon test statistics comparing the distribution of Q_{ih} scores for sets of randomly relabeled NHGRI-GWAS/non-GWAS SNPs. The Wilcoxon test statistic for the true labeling is in red. The permutation p-value associated with the Wilcoxon test is $P < 10^{-5}$ given 10^5 permutations.